基于CRFs的冶金领域中文专利术语抽取研究*

王密平　王　昊　邓三鸿　吴志祥
(南京大学信息管理学院　南京 210023)
(江苏省数据工程与知识服务重点实验室　南京 210023)

摘要: 【目的】探讨冶金领域中文专利术语抽取模型的最优条件, 用于有效地抽取冶金领域专利术语。【方法】使用尚不完善的核心语料库, 在无需人工标引的情况下, 采用条件随机场(CRFs)构建字角色标注的冶金领域中文专利术语识别模型。详细说明模型的构建过程, 同时重点对比CFRs的各个因素(特征组合、字长窗口等)对识别效果的影响。【结果】实验结果表明字序列、级别特征、领域特征、温度特征的组合在字长窗口为3, c等于1, f等于1时, 准确率达到94.26%, 召回率达到94.37%, F1值达到94.5%。【局限】核心词典欠完善, 使得部分词语标注不够准确; 未与其他方法作详细比较, 未详细说明CRFs的可靠性。【结论】CRFs在适当的角色和特征以及特征模板的组合下能较好地识别出冶金领域的中文专利术语。

关键词: 中文专利术语　条件随机场　术语抽取　序列标注
分类号: TP393　G350

1 引　言

专利具有新颖、实用的特征, 是科技信息最为有效的载体之一, 代表了一个国家一个民族的科技发展水平[1]。专利的有效利用能够提高国家和企业的发展速度[2-3]。然而由于中文专利文献为非结构化文本, 并且其中的专利术语包含较多的长术语和英文缩写术语, 作为专利文献核心内容的专利术语较难被科技人员直接识别, 进而影响专利的利用率。因此, 专利术语的抽取显得较为重要。不仅如此, 专利术语的抽取也为分词、句法分析、专利本体的构建等奠定了基础。

目前中文领域术语识别主要有三种方法:

(1) 基于规则的方法[4-5]。基于规则的方法也可称为基于语言学的方法, 主要是根据语言学知识制定特殊句法结构或模板, 匹配符合这些特征模板的字符串。由于特定语言的复杂性, 及其语法不断发展变化, 随着科技的发展, 新术语层出不穷, 使得该方法较难实施, 缺乏灵活性。

(2) 基于统计的方法[6-7]。该方法以统计学为理论基础, 利用语料库中已有的术语分布统计来识别术语。常用的统计方法分为衡量词或词组的领域性, 如词频(Frequency)和衡量词组的单元性, 如互信息(Mutual Information)[8]。

(3) 规则与统计相结合的方法[9-10]。此方法可在统计处理之后采用语法过滤器, 抽取符合统计意义且与给定词法模板匹配的词汇, 也可采用语法规则筛选出候选项, 再计算候选项的统计意义[6]。

条件随机场(Conditional Random Fields, CRFs)是一种典型的序列标注判别模型, 它是在给定的观察序列的条件下, 计算整个观察序列状态标记的联合条件概率分布的无向图模型。CRFs 在隐马尔科夫模型

通讯作者: 王密平, ORCID: 0000-0002-6089-2499, E-mail: 15251890786@163.com。

*本文系江苏省自然科学基金项目“面向专利预警的中文本体学习研究”(项目编号:BK20130587)和江苏省“333”工程项目“面向知识服务的中文本体学习研究”(项目编号:BRA2015401)的研究成果之一。

(Hidden Markov Model, HMM)[11-12]和最大熵模型(Max Entropy Model, MEM)[13-15]基础上建立，克服了HMM[16]以及 MEM 的一些缺点，如 CRFs 对整个标记序列计算联合概率，在整个序列范围内归一化，避免了 MEM 因求解单个或局部观察值概率所带来的标记偏置问题[17]。CRFs 被广泛应用于中文文本的处理。例如邓三鸿等将其用于中文书目关键词标引，论证了该模型的合理性和实用性[18]。王昊等将其应用于网络舆情分析中的人名识别，验证了 CRFs 优于 HMM，探讨了 CRFs 识别人名的最佳条件[19]。刘伙玉等将其用于段落自动划分与构成要素识别，认为 CRFs 在更大的时间复杂度代价下处理效果优于 MEM[20]。将 CRFs 用于专利术语的研究较少，如李鹏等在条件随机场的基础上提出基于规则的摘要信息抽取方法，但其准确率、召回率、F1 值均在 50% 以下，并且规则的手工编写费时费力[21]。刘辉等通过制定语料标注规则进行人工标注，同时采用基于字的序列标注，用 CRFs 进行训练和测试，实现了通信领域的术语抽取，最高准确率为 80%，但其中的人工标注规则仍是一个耗时长的工程，而且未讨论利用术语识别的特征和角色，不利于后人进行更大规模的术语抽取[22]。黄绍杉等使用 CRFs 处理专利的英文摘要，抽取摘要中表示技术和功效内容的信息，平均准确率约为 40%[23]。李洪政等基于 CRFs 识别汉语专利文本介词短语，准确率达 90%以上[24]，但主要通过词性标注角色和特征，在语言学的角度识别介词，实践应用较少。

本文以钢铁冶金领域的中文专利文献的标题为语料，通过核心语料库自动标引字角色以及特征，采用 CRFs 模型，建立中文冶金领域术语自动抽取模型，并通过调整不同的实验参数，观察不同的实验效果来探讨模型最佳识别条件。

2　基于 CRFs 的字角色专利术语识别模型构建

模型分为三部分：字角色的定义、特征标注和角色标注、构建特征模板。字角色标注和特征标注为文本标注，角色标注重点依赖于核心词汇库，用于识别术语的映射和还原。而特征的选择依赖于特定的语料，用于辅助术语识别。特征模板用于控制特征的个数、字长窗口等因素。三者共同形成 CRFs 输入要素。

2.1　字角色标注模型

专利术语识别模型如图 1 所示，整体分成语料生成和序列标注两大部分。在语料生成部分：首先构建钢铁冶金领域中文核心词汇库，包括领域词汇列表，以及常用化学元素等，这些词汇来源于网站、专业词典、专利常用词以及领域专家，共计 6 467 个；然后将专利文本题名进行文本标注，将题名拆分成字序列，包括汉字和连续字母或数字串；通过字角色空间模型标注相应的角色；将字序列和角色序列组合，形成包含字与角色的学习语料。而在序列标注部分，融入外部特征以有效扩展观察序列。首先将专利文本的外部特征，如是否是音译字、是否是姓氏等特征，扩展到学习语料生成观察标注序列构成训练语料；然后训练语料结合特征模板通过 CRFs 算法计算生成序列标注模型，此处将会多次测

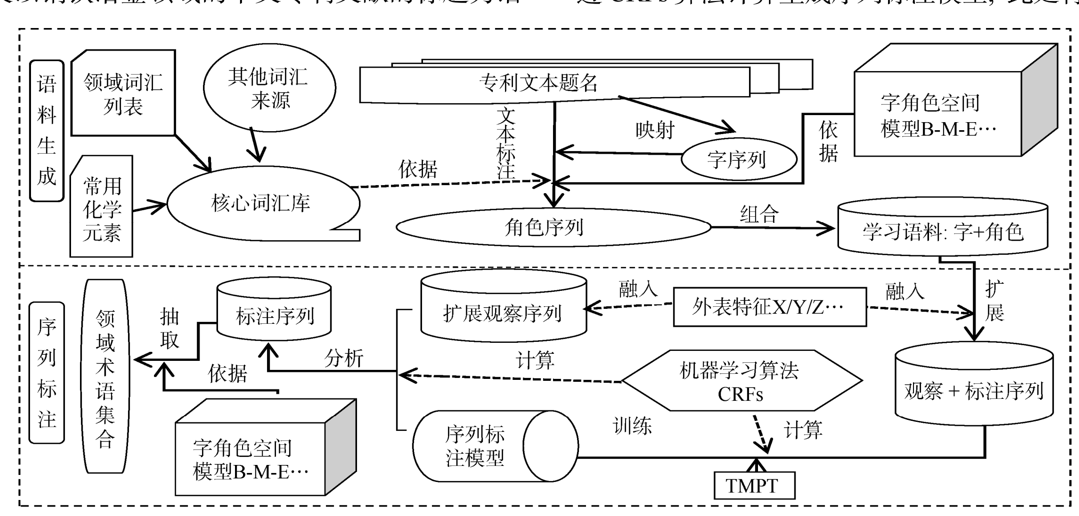


图 1　基于 CRFs 的字角色专利术语识别模型

试不同的观察序列取值，不同的特征集合个数以及不同的角色集合，以寻找最佳的建模条件；仅有观察序列的测试语料在训练的模型学习下生成角色序列；最后依据之前定义的字角色抽取领域术语。

训练和测试语料来自中国国家知识产权局专利检索平台，笔者下载了与该领域相关的中文专利文献共计 7 597 条，并以其题名作为术语抽取的实验文本，前 1 000 条作为测试语料，其余为训练语料。采用的实验工具为 CRF++0.58。

2.2 专利术语角色和特征的定义及其标注

字角色即为观察对象的标注记号，特征是对字序列特征的扩展，字序列与扩展特征共同决定了字所表现出的角色。在角色的定义和标注阶段，首先定义角色类型，其次定义特征类型，接着将字序列扩展标注角色序列和特征序列。

(1) 字角色空间模型的定义

字角色的作用表现在两方面：在语料生成阶段，字序列根据核心词汇库标注字角色；在序列标注阶段，一方面将会影响标注模型的生成，另一方面在最后抽取阶段，需要根据角色集合映射还原成术语，会直接影响到术语抽取的准确率。笔者最终定义了 8 种角色，如表 1 所示。

表 1 字角色集合

角色(R)	说明	示例
B	术语首字	如“脱氧剂”之“脱”
M	术语中字	如“脱氧剂”之“氧”
E	术语尾字	如“脱氧剂”之“剂”
P	术语首字的前一个字	如“一种炼钢生产的新型脱氧剂”之“型”
S	单字术语	如“铁”、“锰”等
A	非术语词中的字	如“一种钢水脱氧剂”之“一”“种”
T	符号数字串	如“GCr15 轧辊淬火加工工艺”之“GCr15”
Q	术语尾字的后一个字	如“一种新硅钙钡镁复合脱氧剂及其生产方法”之“及”

(2) 特征的定义

特征的作用在于扩展语境特征，提高测试阶段的准确率，它依赖于特定的语料。笔者通过观察来源语料发现冶金领域文本的一些特点：冶金术语中包含的化学元素较多，如铝、铁、锰等，并且其中一些为字符，比如化学元素的英文字符 Fe，Q235，NbCFe-Mn-Si 等；范畴词较多，例如，工艺、装置、设备、系统等；关于温度的词语，例如火、热、冷等出现的频率较高。由此，总结得到的特征定义如表 2 所示：

表 2 观察+标注序列标记、取值、描述及示例

观察序列	取值情况	描述
字序列(Z)	汉字或连续字符串	字形特征
姓氏特征(X)	Y	姓氏字(505)
	N	非姓氏字
音译特征(Y)	Y	音译外来字(566)
	N	非音译外来字
领域特征(K)	Y	领域常用字(559)
	N	非领域常用字
级别特征(G)	X	一级常用字(2500)
	Y	二级常用字(1000)
	Z	其他
分类特征(C)	X	指事字(184)
	Y	象形字(244)
	Z	形声字(3505)
	U	会意字(673)
	V	其他类型字
温度特征(T)	Y	温度词(76)
	N	非温度词

示例

Z	X	Y	K	G	C	T	R
一	N	Y	Y	X	X	N	A
种	N	N	N	Y	Z	N	P
燃	N	N	Y	X	Z	Y	B
气	N	N	Y	X	U	N	E
加	N	Y	Y	X	Y	Y	B
热	N	Y	Y	X	Z	Y	E
大	N	Y	Y	X	U	N	Q
型	N	N	Y	X	Z	N	A
连	Y	Y	Y	X	Y	N	A
续	N	N	Y	X	Z	N	A
式	N	N	Y	X	Z	N	A
网	N	N	Y	X	U	N	A
带	N	N	Y	X	U	N	P
淬	N	N	Y	Z	Z	Y	B
火	Y	N	Y	X	U	Y	M
炉	N	N	Y	X	Z	Y	E

在表 2 的示例部分，字序列(Z)中竖向的虚线表示纵向序列组合约束，常用的包括远程上下文信息和局部上下文信息，前者指与当前对象具有一定文本距离的对象所提供的长距离约束，后者指以当前汉字为中心，向前或(和)向后连续选取一定长度范围的上下文作为当前汉字的约束，这个局部连续范围称为字长窗口[25]，常用的有 3 字长窗口和 5 字长窗口。该示例为 5 字长窗口，后续实验中将详细比较 3 字长窗口和 5 字长窗口对结果的影响。横向的虚线为横向序列组合约束。本文将在语料中出现的连续阿拉伯数字和连续的英文字母作为一个单字处理。

(3) 字序列、角色序列、扩展序列的生成

角色标注的算法首先将句子拆分为单个字符，并且将连续的字符或数字合并为一个整体。存入二维数组的第一列构成字序列。然后判断句子中是否包含核心词汇，如果包含则标记句子中核心词汇的角色，并且依次映射到字序列。最后标注包含数字或字母的字符串，以及非术语。扩展序列根据标题中的单个字符 R 是否在相应的语料中，来标注相应的特征符号。

2.3　特征模板的构建

特征模板描述了在训练和测试阶段中用到的特征。模板文件中每一行代表一个模板，在每个模板里，特定的宏%x[Row, Col]用来描述输入数据的片段，Row 表示当前片段的相对位置，Col 则表示列的绝对位置。在表 3 的特征模板中 n 为 Row，取值为 0 代表当前位置，–1 表示当前字的上一个字，1 为当前字的下一个字。而 n-gram 表示多元特征关系，如 1-gram 表示 1-元关系特征，2-gram 表示 2-元关系特征。为探讨不同特征组合的识别效果，笔者设置了 10 个模板，如表 3 所示：

表 3　冶金术语角色标注的特征模板

模板名称	观察特征	标注角色	n-gram	特征模板
TMPT0	Z	L	1-gram	Z_n, n=-2, -1, 0, 1, 2
			2-gram	$Z_{n-1}Z_n$, n=-1, 0, 1, 2; $Z_{n-2}Z_n$, n=0, 1, 2; $L_{-1}L_0$
			3-gram	$Z_{n-2}Z_{n-1}Z_n$, n=0, 1, 2
TMPT1	ZX	L	1-gram	Z_n, X_n, Z_nX_n, n=-2, -1, 0, 1, 2
			2-gram	$Z_{n-1}Z_n$, $X_{n-1}X_n$, n=-1, 0, 1, 2;$Z_{n-2}Z_n$, $X_{n-2}X_n$, n=0, 1, 2; $L_{-1}L_0$
			3-gram	$Z_{n-2}Z_{n-1}Z_n$, $X_{n-2}X_{n-1}X_n$, n=0, 1, 2
TMPT2	ZXY	L	1-gram	Z_n, X_n, Y_n, $Z_nX_nY_n$, n=-2, -1, 0, 1, 2
			2-gram	$Z_{n-1}Z_n$, $X_{n-1}X_n$, $Y_{n-1}Y_n$, n=-1, 0, 1, 2; $Z_{n-2}Z_n$, $X_{n-2}X_n$, $Y_{n-2}Y_n$, n=0, 1, 2; $L_{-1}L_0$
			3-gram	$Z_{n-2}Z_{n-1}Z_n$, $X_{n-2}X_{n-1}X_n$, $Y_{n-2}Y_{n-1}Y_n$, n=0, 1, 2
TMPT3	ZXYK	L	1-gram	Z_n, X_n, Y_n, K_n, $Z_nX_nY_nK_n$, n=-2, -1, 0, 1, 2
			2-gram	$Z_{n-1}Z_n$, $X_{n-1}X_n$, $Y_{n-1}Y_n$, $K_{n-1}K_n$, n=-1, 0, 1, 2; $Z_{n-2}Z_n$, $X_{n-2}X_n$, $Y_{n-2}Y_n$, $K_{n-2}K_n$, n=0, 1, 2; $L_{-1}L_0$
			3-gram	$Z_{n-2}Z_{n-1}Z_n$, $X_{n-2}X_{n-1}X_n$, $Y_{n-2}Y_{n-1}Y_n$, $K_{n-2}K_{n-1}K_n$, n=0, 1, 2
TMPT4	ZXYKG	L	1-gram	Z_n, X_n, Y_n, K_n, G_n, $Z_nX_nY_nK_nG_n$, n=-2, -1, 0, 1, 2
			2-gram	$Z_{n-1}Z_n$, $X_{n-1}X_n$, $Y_{n-1}Y_n$, $K_{n-1}K_n$, $G_{n-1}G_n$, n=-1, 0, 1, 2; $Z_{n-2}Z_n$, $X_{n-2}X_n$, $Y_{n-2}Y_n$, $K_{n-2}K_n$, $G_{n-2}G_n$, n=0, 1, 2; $L_{-1}L_0$
			3-gram	$Z_{n-2}Z_{n-1}Z_n$, $X_{n-2}X_{n-1}X_n$, $Y_{n-2}Y_{n-1}Y_n$, $K_{n-2}K_{n-1}K_n$, $G_{n-2}G_{n-1}G_n$, n=0, 1, 2
TMPT5	ZXYKGC	L	1-gram	Z_n, X_n, Y_n, K_n, G_n, C_n, $Z_nX_nY_nK_nG_nC_n$, n=-2, -1, 0, 1, 2
			2-gram	$Z_{n-1}Z_n$, $X_{n-1}X_n$, $Y_{n-1}Y_n$, $K_{n-1}K_n$, $G_{n-1}G_n$, $C_{n-1}C_n$, n=-1, 0, 1, 2; $Z_{n-2}Z_n$, $X_{n-2}X_n$, $Y_{n-2}Y_n$, $K_{n-2}K_n$, $G_{n-2}G_n$, $C_{n-2}C_n$, n=0, 1, 2; $L_{-1}L_0$
			3-gram	$Z_{n-2}Z_{n-1}Z_n$, $X_{n-2}X_{n-1}X_n$, $Y_{n-2}Y_{n-1}Y_n$, $K_{n-2}K_{n-1}K_n$, $G_{n-2}G_{n-1}G_n$, $C_{n-2}C_{n-1}C_n$, n=0, 1, 2
TMPT6	ZXYKGC	L	1-gram	Z_n, X_n, Y_n, K_n, G_n, C_n, $X_nY_nK_nG_nC_n$, n=-1, 0, 1
			2-gram	$Z_{n-1}Z_n$, $X_{n-1}X_n$, $Y_{n-1}Y_n$, $K_{n-1}K_n$, $G_{n-1}G_n$, $C_{n-1}C_n$, n=0, 1; $Z_{n-2}Z_n$, $X_{n-2}X_n$, $Y_{n-2}Y_n$, $K_{n-2}K_n$, $G_{n-2}G_n$, $C_{n-2}C_n$, n=1; $L_{-1}L_0$
			3-gram	$Z_{n-2}Z_{n-1}Z_n$, $X_{n-2}X_{n-1}X_n$, $Y_{n-2}Y_{n-1}Y_n$, $K_{n-2}K_{n-1}K_n$, $G_{n-2}G_{n-1}G_n$, $C_{n-2}C_{n-1}C_n$, n=1
TMPT7	ZXYKGC	L		同 TMP5，仅除去 $L_{-1}L_0$
TMPT8	ZXYKGCT	L	1-gram	Z_n, X_n, Y_n, K_n, G_n, C_n, T_n, $Z_nX_nY_nK_nG_nC_nT_n$, n=-2, -1, 0, 1, 2
			2-gram	$Z_{n-1}Z_n$, $X_{n-1}X_n$, $Y_{n-1}Y_n$, $K_{n-1}K_n$, $G_{n-1}G_n$, $C_{n-1}C_n$, $T_{n-1}T_n$, n=-1, 0, 1, 2
			3-gram	$Z_{n-2}Z_{n-1}Z_n$, $X_{n-2}X_{n-1}X_n$, $Y_{n-2}Y_{n-1}Y_n$, $K_{n-2}K_{n-1}K_n$, $G_{n-2}G_{n-1}G_n$, $C_{n-2}C_{n-1}C_n$, $T_{n-2}T_{n-1}T_n$, n=0, 1, 2
TMPT9	ZXYKGC	L		同 TMPT5 除去 $Z_{n-2}Z_n$, $X_{n-2}X_n$, $Y_{n-2}Y_n$, $K_{n-2}K_n$, $G_{n-2}G_n$, $C_{n-2}C_n$, n=0, 1, 2; $L_{-1}L_0$

TMPT0, TMPT1, TMPT2, TMPT3, TMPT4, TMPT5, TMPT6 依次扩展特征。TMPT5 和 TMPT6 用于比较 3 字长窗口与 5 字长窗口的差别。TMPT4 和 TMPT7 用于探讨上一个字角色对当前字角色的约束对结果的影响程度。

3 专利术语字角色标注模型实验分析

经过实验以后, 在抽取阶段, 通过字角色空间映射还原成术语。如表 1 里角色定义的一样, B 为术语首字, E 为尾字, S 为单字术语。那么表 2 中 BE 燃气为一个术语, BME 淬火炉为一个术语。“种”为术语“燃气”的前一个字, 而“大”为术语“加热”的后一个字。本文约定, 识别出的正确术语为抽取后的领域集合中的单字术语以及完整非单字术语即以 B 开头, 以 E 结尾中间为 M 的术语。识别出的术语为所有单字术语以及以 B 开头的术语。所有标注的术语为核心词汇库中的术语。

本文采用以下指标衡量实验结果: 准确率 P、召回率 R、F1 值、以及单字识别率 SP。

$$P=\frac{\text{识别出的正确术语个数(RN)}}{\text{识别出的术语个数(STN)}}\times 100\%$$

$$R=\frac{\text{识别出的正确术语个数(RN)}}{\text{所有标注的术语个数(TN)}}\times 100\%$$

$$F1=\frac{2PR}{P+R}$$

$$SP=\frac{\text{识别出的字个数}}{\text{所有的字个数}}$$

3.1 不同特征模板对比

根据表 3 设置的特征模板, 得到的结果如表 4 和图 2 所示。由于各模板的单字识别率均较高, 基本在 94.5%以上, 比较意义不大, 故未在图 2 中列出。鉴于篇幅限制, 未列出正确识别数(RN)、识别数(STN)、以及标注的所有术语数(TN)随着模板的变化情况, 详细数据如表 4 所示:

表 4 不同特征模板的专利术语识别结果

指标＼模板	TMPT0	TMPT1	TMPT2	TMPT3	TMPT4	TMPT5	TMPT6	TMPT7	TMPT8	TMPT9
P	93.14%	92.17%	92.07%	92.47%	93.43%	93.10%	93.17%	91.32%	93.18%	93.29%
R	92.49%	92.26%	92.16%	92.51%	93.78%	93.29%	93.68%	91.34%	93.83%	93.63%
F1	92.81%	92.22%	92.12%	92.49%	93.61%	93.19%	93.43%	91.33%	93.51%	93.46%
SP	95.11%	94.58%	94.50%	94.63%	95.00%	94.65%	94.90%	94.43%	95.09%	94.96%
RN	3 705	3 696	3 692	3 706	3 757	3 737	3 753	3 659	3 759	3 751
STN	3 978	4 010	4 010	4 008	4 021	4 014	4 028	4 007	4 034	4 021
TN	4 006	4 006	4 006	4 006	4 006	4 006	4 006	4 006	4 006	4 006

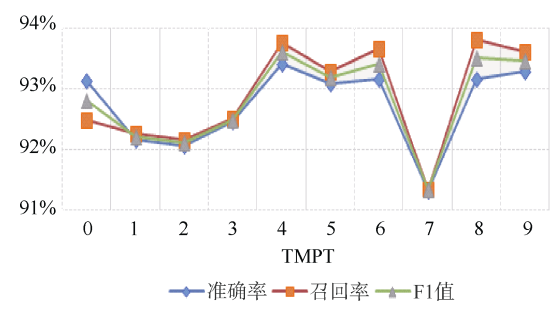


图 2 不同特征模板的比较结果

(1) 特征增加的作用探讨

TMPT0, TMPT1, TMPT2, TMPT3, TMPT4, TMPT5, TMPT8 用于对比特征的增加对实验效果的影响。这几个模板均使用 5 字长窗口。如图 2 所示, 在 TMPT0 测试时, 即只有字序列与角色两列时, 所得准确率达到 93.14%, 召回率达到 92.49%, 这说明字本身占据了主导作用。随着特征的扩展, 各指标稍有下降, 随后上升, 直至 TMPT4 增加级别特征后, 正确识别出的术语个数多达 3 757 个, 准确率达到 93.43%, 召回率达到 93.78%, F1 值也达到 93.61%, 同时单字识别率也达到了最大值 95.00%。这说明合适的特征扩展能够提高识别率, 而不相关的特征反而会干扰术语的识别。同时从整体变化趋势看, 召回率的变化比准确率的变化大, 说明了特征的增加更有利于术语的召回。

(2) 纵向、横向的制约作用以及前一角色对当前角色的制约作用

TMPT5 和 TMPT6 探讨纵向约束的作用。TMPT5 为 5 字长窗口, 而 TMPT6 为 3 字长窗口, 其比较结果如图 2 示, 两者结果差异不大, 但是总体上 TMPT6 的准确率、召回率和 F1 值略高。这说明字长窗口的增加并不与识别效果成正比, 需要视具体语料而定。如果选择不当, 在一定程度上会降低识别效果。

TMPT5 和 TMPT9 探讨横向间隔的特征之间的影响。TMPT9 去除了横向间隔特征之间的约束。两者结果差异较小，但整体上 TMPT9 的各指标值略高，说明间隔特征的约束不一定能提高识别效果。在当前语料中更适宜去除间隔特征的影响。

TMPT5 和 TMPT7 比较前一个字对当前字的制约作用。TMPT7 去除了前一角色对当前角色的约束，结果显示其各个指标比 TMPT5 明显降低，这说明前一角色对当前角色的约束作用非常重要。

(3) 不同特征组合结果的变化

为探讨不同特征的作用，笔者结合上一步的实验结果，重新调整特征模板，使用 3 字长窗口，同时去除间隔特征的约束，并且保留 $L_{-1}L_0$，将有用特征重新组合实验验证识别效果。

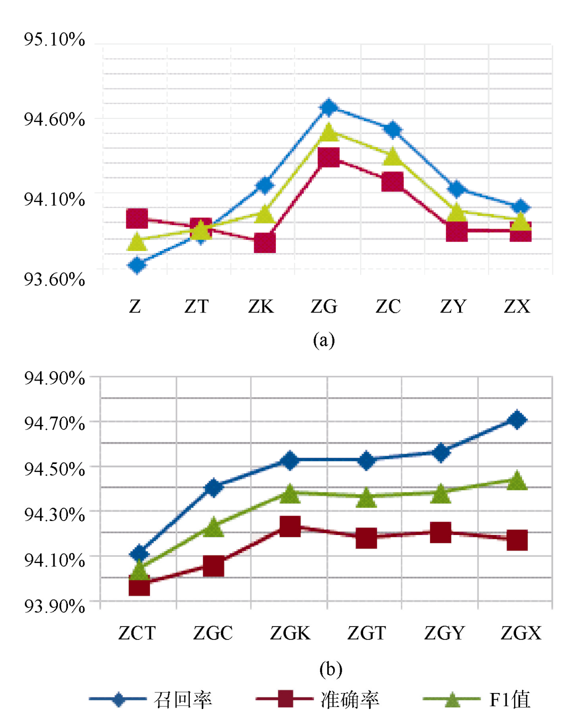


图 3　单个特征比较和两个特征比较

①单个特征与字序列组合，探讨较为有用的特征。鉴于篇幅，笔者未列出数据表格，仅绘制趋势图，如图 3(a)所示。整体上看任何特征的增加，召回率均有所提高。而准确率与初始情况对比，有降低现象。从单个增加的特征对比来看，特征 G，即级别特征最有利于识别专利术语。其次是 C 分类特征，而领域特征 K、音译特征 Y 效果相似，姓氏特征 X 和温度特征 T 效果最差。

②笔者选择效果较好特征 G、特征 C 与其他特征组合探讨组合效果。结果如图 3(b)所示。将 C 和 T 组合效果作为基准对比，然后将效果最佳的 G 与其他特征组合，结果发现 G 和 C 组合并未达到最佳状态，相反 G 与其他特征组合效果更好，同时 G 与 X 组合时召回率达到最大值 94.71%，与 K 组合时准确率达到了最大值 94.23%。这说明特征的组合并不与单个特征的效果成正相关。

3.2　角色定义增加

前面的实验中，角色定义均为 B、M、E、A、S 共 5 个角色。为探讨角色定义的增加对实验结果的影响，笔者新增了两个角色 P、Q，P 表示术语首字的前一个字，而 Q 表示术语尾字的后一个字，详细示例可参见表 1。当术语为连续术语时，如表 2 中的“燃气”和“加热”两个术语，此时术语的后一个字还是术语，优先标注术语，只有当术语的前或后一个字为非术语时，将其标注为 P 或 Q。经过对特征模板的实验筛选，选择改进后的 6 个模板，以及相同特征组合顺序，测试结果如图 4 所示，在只有字序列本身时，两者差距较小，随着特征的增加，无 P 和 Q 角色的识别效果较好。这说明角色的定义需适当，角色不恰当的增加反而不利于识别。

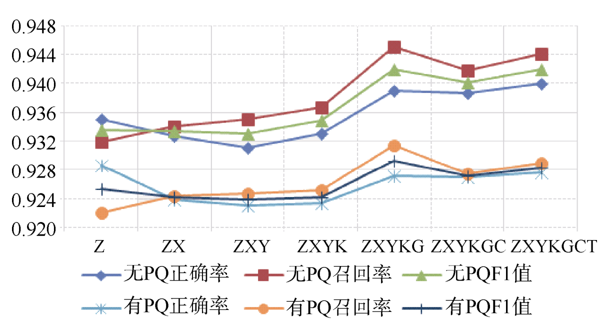


图 4　增加角色 P 和角色 Q 后的识别效果

3.3　不同参数对比

观察以上特征以及角色组合，选择 P 值、R 值和 F1 值均最高的模板，即由 ZTKG 组成的模板进行软件边界参数 c 值，以及特征函数频次阈值 f 值的调整实验。c 用于调节条件随机场模型中的数据欠拟合和过拟合之间的平衡。f 用于限制训练数量中出现不少于 f 次的特征。由图 5(a)可知，当 f 值为 1 时，识别效果最好，随着 f 值的增加，准确率、召回率、F1 值均下降。这可能与本文所使用的专利文献语料特征较少有关。随着 f 值的增加，低频的特征被过滤，导致识别出的正确术语数量减少。图 5(b)中显示，c 值的变化对识别效果整体波动不大，从 1 增大到 4 时，呈现上升趋势，随后迅速下降，而后又逐渐上升，到 c 等于 9 时，各项指标达到最大值。这说明 c 值的变化对识别效果整体影响不大。

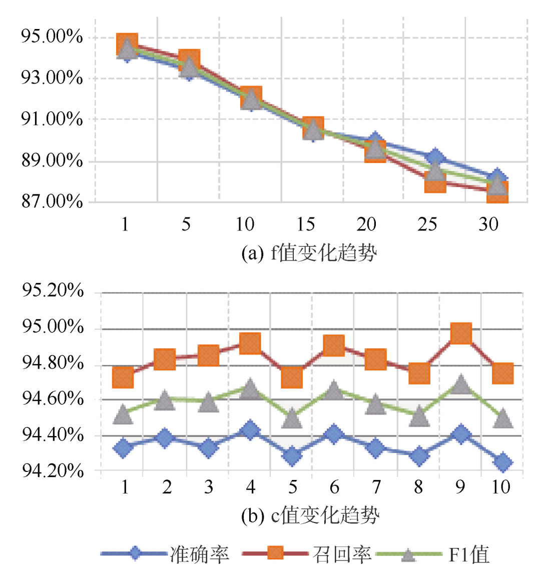


图 5 频率参数 f 值调整和拟合参数 c 值调整

4 结 语

本文通过定义不同的角色和特征，同时对原始专利术语语料进行序列标注生成学习语料和测试语料，通过 CRFs 模型抽取术语。其中详细探讨了特征模型、角色和参数三个因素对结果的影响，实验结果表明：

(1) 恰当的扩展特征序列有助于术语识别，否则特征的增加反而不利于术语识别；二元特征的约束明显有助于术语识别；间隔特征的约束在本文语料中不利于术语识别。

(2) 角色的增加不一定与识别效果成正比，需要根据具体语料进行调整。

(3) c 值变化整体而言对实验结果影响不大，在特征较少的专利文献中，f 值为 1 时识别效果最好。

(4) 以不完善的核心语料库作为原始标引语料的前提下，7 597条冶金领域的题名学习和训练时间约85.45s，CRFs 在效果最优的角色、特征、以及特征模板的实验中，得到94%以上的准确率和召回率，同时获取到正确的未登录词共70个，例如“预热器”、“卤化物”、“电炉炼钢炉”、“反应剂”、“锻热”、“均热钢”、“模具炉”等。其准确率高并且可识别一定数量的未登录词，说明该模型优于 HMM 等基于规则的识别方法。

但文中也存在不可避免的误差因素，例如以核心词汇库代替人工标引的学习语料库节省了标注时间，但会产生语料标注不充分问题。其次该语料为冶金领域专利术语的标题，相对于正文而言，更为精炼和整齐，使得准确率、召回率、F1 值较高。由于测试条件的限制，未能训练更大的样品进行实验。今后可根据以上实验结果直接设置较为有用的特征组合，设置最为有效的特征模板以及参数，进行摘要和正文的实验，同时邀请专家对未登录词进行判断，以期在最小耗时和最小专家成本下最大限度地识别出更多正确术语。

参考文献：

[1] 贺延芳. 专利文献研究助力我国创新活动[N]. 中国知识产权报, 2012-03-23(4). (He Yanfang. The Patent Literature Study Assist in Chinese Innovation Activities [N]. China Intellectual Property News, 2012-03-23(4).)

[2] 葛煦，卢宝华，杨湘华，等. 谈高校科技发展中专利文献的利用[J]. 技术与创新管理, 2005, 26(1): 68-70. (Ge Xu, Lu Baohua, Yang Xianghua, et al. Utilization of Patent Literature on the Development of Science and Technology in Universities [J]. Technology and Innovation Management, 2005, 26(1): 68-70.)

[3] 贾志琦，邵曰剑. 有效利用专利文献提高企业技术创新能力[J]. 山西科技, 2008(1): 91-93. (Jia Zhiqi, Shao Yuejian. Enhance Enterprises' Technological Innovative Capability Through Effective Use of Patent Documents [J]. Shanxi Science and Technology, 2008(1): 91-93.)

[4] Uzunbas M G, Chen C, Metaxas D. An Efficient Conditional Random Field Approach for Automatic and Interactive Neuron Segmentation [J]. Medical Image Analysis, 2016, 27: 31-44.

[5] 张雷瀚，吕学强，李卓，等. 领域本体术语的抽取方法研究[J]. 情报学报, 2014, 33(2): 167-174. (Zhang Leihan, Lv Xueqiang, Li Zhuo, et al. Research on Extraction Methods for Domain Ontology Terminology [J]. Journal of the China Society for Scientific and Technical Information, 2014, 33(2): 167-174.)

[6] 袁劲松，张小明，李舟军，等. 术语自动抽取方法研究综述[J]. 计算机科学, 2015, 42(8): 7-12. (Yuan Jinsong, Zhang Xiaoming, Li Zhoujun, et al. Survey of Automatic Terminology Extraction Methodologies [J]. Computer Science, 2015, 42(8): 7-12.)

[7] 汤青，吕学强，李卓，等. 领域本体术语抽取研究[J]. 现代图书情报技术, 2014(1): 43-50. (Tang Qing, Lv Xueqiang, Li Zhuo, et al. Research on Domain Ontology Term Extraction [J]. New Technology of Library and Information Service,

2014(1): 43-50.)

[8] 王昊, 刘建华, 苏新宁, 等. 面向语义网的本体学习技术和系统研究[J]. 现代图书情报技术, 2009(1): 64-72. (Wang Hao, Liu Jianhua, Su Xinning, et al. Research on Techniques and Systems of Ontology Learning for Semantic Web [J]. New Technology of Library and Information Service , 2009(1): 64-72.)

[9] 谷俊, 王昊. 基于领域中文文本的术语抽取方法研究[J]. 现代图书情报技术, 2011(4): 29-34. (Gu Jun, Wang Hao. Study on Term Extraction on the Basis of Chinese Domain Texts [J]. New Technology of Library and Information Service, 2011(4): 29-34.)

[10] 化柏林. 针对中文学术文献的情报方法术语抽取[J]. 现代图书情报技术, 2013(6): 68-75. (Hua Bolin. Extracting Information Method Term from Chinese Academic Literature [J]. New Technology of Library and Information Service, 2013(6): 68-75.)

[11] Zhou H T, Chen J, Dong G M, et al. Detection and Diagnosis of Bearing Faults Using Shift-invariant Dictionary Learning and Hidden Markov Model [J]. Mechanical Systems and Signal Processing, 2016, 72-73: 65-79.

[12] 乐娟, 赵玺. 基于 HMM 的京剧机构命名实体识别算法[J]. 计算机工程, 2013, 39(6): 266-271, 286. (Le Juan, Zhao Xi. Algorithm of Beijing Opera Organization Names Entity Recognition Based on HMM [J]. Computer Engineering, 2013, 39(6): 266-271, 286.)

[13] 李丽双, 王意文, 黄德根. 基于信息熵和词频分布变化的术语抽取研究[J]. 中文信息学报, 2015, 29(1): 82-87. (Li Lishuang, Wang Yiwen, Huang Degen. Term Extraction Based on Information Entropy and Word Frequency Distribution Variety [J]. Journal of Chinese Information Processing, 2015, 29(1): 82-87.)

[14] 卢达威, 宋柔. 基于最大熵模型的汉语标点句缺失话题自动识别初探[J]. 计算机工程与科学, 2015, 37(12): 2282-2293. (Lu Dawei, Song Rou. Automatic Recognition of the Absent Topics in Chinese Punctuation Clauses Based on Maximum Entropy Model [J]. Computer Engineering and Science, 2015, 37(12): 2282-2293.)

[15] 何径舟, 王厚峰. 基于特征选择和最大熵模型的汉语词义消歧[J]. 软件学报, 2010, 21(6): 1287-1295. (He Jingzhou, Wang Houfeng. Chinese Word Sense Disambiguation Based on Maximum Entropy Model with Feature Selection [J]. Journal of Software, 2010, 21(6): 1287-1295.)

[16] 王昊, 邓三鸿. HMM 和 CRFs 在信息抽取应用中的比较研究[J]. 现代图书情报技术, 2007(12): 57-63. (Wang Hao, Deng Sanhong. Comparative Study on HMM and CRFs Applying in Information Extraction [J]. New Technology of Library and Information Service, 2007(12): 57-63.)

[17] Song D J, Liu W, Zhou T Y et al. Efficient Robust Conditional Random Fields [J]. IEEE Transactions on Image Processing, 2015, 24(10): 3124-3136.

[18] 邓三鸿, 王昊, 秦嘉杭, 等. 基于字角色标注的中文书目关键词标引研究[J]. 中国图书馆学报, 2012, 38(2): 38-49. (Deng Sanhong, Wang Hao, Qin Jiahang, et al. Research on Keywords Indexing for Chinese Bibliography Based on Word Roles Annotation [J]. Journal of Library Science in China, 2012, 38(2): 38-49.)

[19] 王昊, 苏新宁. 基于CRFs的角色标注人名识别模型在网络舆情分析中的应用[J]. 情报学报, 2009, 28(1): 88-96. (Wang Hao, Su Xinning. Model for Person Name Recognition Based on Role Labeling Using CRFs and Its Application to Web Opinion Analysis [J]. Journal of the China Society for Scientific and Technical Information, 2009, 28(1): 88-96.)

[20] 刘伙玉, 王东波, 苏新宁. 多特征下的科研论文段落自动划分与构成要素识别研究[J]. 情报学报, 2015, 34(4): 388-397. (Liu Huoyu, Wang Dongbo, Su Xinning. Research of Paragraphs Segmentation and Elements Recognition for Academic Papers Based on Multi-features [J]. Journal of the China Society for Scientific and Technical Information, 2015, 34(4): 388-397.)

[21] 李鹏, 桂婕, 乔晓东, 等. 条件随机场与规则集成的专利摘要信息抽取[J]. 数字图书馆论坛, 2010(9): 2-6. (Li Peng, Gui Jie, Qiao Xiaodong, et al. Patent Summary Information Extraction Based on Conditional Random Fields and Rule Integrated [J]. Digital Library Forum, 2010(9): 2-6.)

[22] 刘辉, 刘耀. 基于条件随机场的专利术语抽取[J]. 数字图书馆论坛, 2014(12): 46-49. (Liu Hui, Liu Yao. Patent Term Extraction Based on Conditional Random Field [J]. Digital Library Forum, 2014(12): 46-49.)

[23] 黄绍杉, 乔晓东, 桂婕, 等. 基于条件随机场的专利摘要信息抽取研究[J]. 数字图书馆论坛, 2010(9): 7-12. (Huang Shaoshan, Qiao Xiaodong, Gui Jie, et al. Research on Summary of Patent Information Extraction Based on Conditional Random Field [J]. Digital Library Forum, 2010(9): 7-12.)

[24] 李洪政, 晋耀红. 基于条件随机场方法的汉语专利文本介词短语识别[J]. 现代语文(语言研究), 2015(7): 120-122. (Li Hongzheng, Jin Yaohong. Recognition of Chinese Patent Text

Prepositional Phrase Based on conditional Random Field [J]. Modern Chinese, 2015(7): 120-122.)

[25] Peng F, McCallum A. Infomation Extraction from Research Papers Using Conditional Random Fields [J]. Information Processing and Management, 2006, 42(4): 963-979.

作者贡献声明:

王密平, 王昊: 提出研究思路, 设计研究方案;
王密平: 进行实验; 采集、清洗和分析数据; 论文起草;
邓三鸿, 吴志祥: 论文最终版本修订。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据见期刊网络版 http://www.infotech.ac.cn。
[1] 王密平. train.txt. 训练数据.
[2] 王密平. test.txt. 测试数据.

收稿日期: 2016-03-01
收修改稿日期: 2016-03-28

Extracting Chinese Metallurgy Patent Terms with Conditional Random Fields

Wang Miping　Wang Hao　Deng Sanhong　Wu Zhixiang
(School of Information Management, Nanjing University, Nanjing 210023, China)
(Jiangsu Key Laboratory of Data Engineering and Knowledge Service, Nanjing 210023, China)

Abstract: [**Objective**] This paper proposed a model to extract metallurgy patent terms in Chinese effectively. [**Methods**] We created the model to automatically identify metallurgy patent terminologies in Chinese with the help of conditional random fields(CRFs) technology. This model was tested with an incomplete core corpus. We discussed the development process and then compared the impacts of various CRFs factors to this character-role-labeled model. [**Results**] The new model combined the character sequences, level features, areal features and temperature features of the patent terms. Its precision rate was 94.26%, the recall rate was 94.37%, and the F1 value was 94.5%, while the length of the proximity window and the values of the parameter c and f were 3, 1, and 1 respectively. [**Limitations**] Some of the term labels were not accurate enough due to the incomplete core corpus. We did not compare our model with other methods to discuss the reliability of the CRFs. [**Conclusions**] The CRFs model could effectively identify the metallurgy patent terms in Chinese under appropriate working conditions.

Keywords: Chinese patent terminology　CRFs　Terminology extraction　Sequence labeling